Comparaison du *Lexique-Grammaire* des verbes pleins et de DICOVALENCE : vers une intégration dans le Le*fff*

Laurence Danlos*, Benoît Sagot†
* Lattice - Université Paris 7 - Institut Universitaire de France
2 place Jussieu, case 7003, 75251 Paris Cedex 05, France
laurence.danlos@linguist.jussieu.fr
† Projet Signes - INRIA
Dom. Universitaire, 351 cours de la Libération, 33405 Talence Cedex, France
benoit.sagot@inria.fr

Résumé Cet article compare le *Lexique-Grammaire* des verbes pleins et DICOVALENCE, deux ressources lexicales syntaxiques pour le français développées par des linguistes depuis de nombreuses années. Nous étudions en particulier les divergences et les empiètements des modèles lexicaux sous-jacents. Puis nous présentons le Le*fff*, lexique syntaxique à grande échelle pour le TAL, et son propre modèle lexical. Nous montrons que ce modèle est à même d'intégrer les informations lexicales présentes dans le *Lexique-Grammaire* et dans DICOVALENCE. Nous présentons les résultats des premiers travaux effectués en ce sens, avec pour objectif à terme la constitution d'un lexique syntaxique de référence pour le TAL.

Abstract This paper compares the *Lexicon-Grammar* of full verbs and DICOVALENCE, two syntactic lexical resources for French developed by linguists for numerous years. We focus on differences and overlaps between both underlying lexical models. Then we introduce the Le*fff*, large-coverage syntactic lexicon for NLP, and its own lexical model. We show that this model is able to integrate lexical information present in the *Lexicon-Grammar* and in DICOVALENCE. We describe the results of the first work done in this direction, the long term goal being the consitution of a high-quality syntactic lexicon for NLP.

Mots-clefs : Lexique syntaxique, *Lexique-Grammaire*, DICOVALENCE, Le*fff*

Keywords: Syntactic lexicon, *Lexicon-Grammar*, DICOVALENCE, Le*fff*

1 Introduction

À l'heure actuelle, il existe deux grandes ressources lexicales syntaxiques pour le français développées depuis de nombreuses années dans des laboratoires de linguistique : le *Lexique-Grammaire* (Gross, 1975; Boons *et al.*, 1976a; Boons *et al.*, 1976b; Guillet & Leclère, 1992) et le dictionnaire DICOVALENCE (Van den Eynde & Mertens, 2006). L'objectif de cet article est de comparer ces deux ressources lexicales, afin d'en tirer le meilleur parti, et de l'intégrer dans le Le*fff* (Lexique des Formes Fléchies du Français, (Sagot, 2006; Sagot & Danlos, 2007), lexique syntaxique destiné au TAL et en cours de développement. Nous nous limitons ici aux seules

entrées des verbes pleins, c'est-à-dire ni figés ni supports d'adjectifs ou de noms prédicatifs (le *Lexique-Grammaire* et le Le*fff* comportent des entrées pour les verbes non pleins et des entrées non verbales, mais ceci n'est pas le cas de DICOVALENCE). Les objectifs des lexiques considérés, comme de tout lexique syntaxique, sont de définir, pour chaque lemme verbal donné, ses différents emplois et, pour chacun de ces emplois, son (ou ses) cadres de sous-catégorisation et les informations complémentaires qui s'y rapportent (p.ex. les informations sur le contrôle).

Nous présentons donc un travail de comparaison entre le *Lexique-Grammaire* et DICOVALENCE, ressources présentées brièvement dans les sections 2 et 3, afin de comprendre leurs points communs et leurs divergences et d'aider à leur amélioration mutuelle (Section 4). Puis nous montrons (Section 5) en quoi le modèle lexical utilisé dans le Le*fff* permet de modéliser les informations présentes dans l'une ou l'autre de ces ressources, améliorant ainsi sa précision, sa couverture, et donc la qualité des outils de TAL qui l'utilisent.

2 Introduction au *Lexique-Grammaire*

Dans le *Lexique-Grammaire*, un cadre de sous-catégorisation d´un emploi de verbe plein, qui donne sa valence, est défini par deux critères de base (non hiérarchisés) :

– nombre et nature directe ou indirecte des compléments : le tableau ci-dessous résume les schémas de cadres de sous-catégorisation définis par ce critère.

$N_0\ V$	zéro complément
$N_0\ V\ N_1$ — $N_0\ V\ Prép\ N_1$	un complément
$N_0\ V\ N_1\ N_2$ — $N_0\ V\ N_1\ Prép\ N_2$	deux compléments
$N_0\ V\ N_1\ Prép\ N_2\ Prép\ N_3$ — $N_0\ V\ Prép\ N_1\ Prép\ N_2\ Prép\ N_3$	trois compléments

Ces schémas sont ensuite affinés selon la valeur des prépositions : sont distinguées Prép = *à*, Prép = *de*, Prép = Loc, Prép = *avec* et Prép = autres.

– réalisations du sujet et des compléments ; nous employons le terme de « position » pour désigner un élément N_i, où N_0 correspond au sujet et N_i ($i > 1$) à un complément. Une position peut-être réalisée comme une complétive (notée $Que\ P$), une infinitive (notée $Vinf$) ou un groupe nominal (noté GN). Pour chaque N_i, les distributions suivantes sont celles qui sont le plus souvent retenues : $N_i = Que\ P\ /\ Vinf\ /\ GN$ ou $N_i = Vinf\ /\ GN$ ou $N_i = GN$. À ces distributions s'ajoutent le sujet impersonnel (pléonastique) réalisé par les pronoms *il* ou *ça*, soit $N_0 =$ *ilimp* ou $N_0 =$ *çaimp*.

Les différents cadres de sous-catégorisation des emplois des verbes sont structurés en *Tables*. Chaque table est définie par une *propriété définitoire*. Parmi les propriétés définitoires, on peut distinguer les propriétés de base de celles qui sont occasionnelles :

– Les propriétés définitoires basiques correspondent à l'intersection des deux critères que nous venons d'expliquer. Ainsi la propriété définitoire de la Table 9 est $N_0\ V\ (QueP)_1\ à\ N_2$, ce qui correspond au schéma $N_0\ V\ N_1\ à\ N_2$ en ce qui concerne le nombre et la nature des compléments, et au fait que la réalisation de N_1 est $N_1 = Que\ P\ /\ Vinf\ /\ GN$. Par exemple, le verbe *dire* appartient à la Table 9 parce qu'il est la tête verbale des phrases suivantes : *(Luc)*$_0$ *a dit à (Marie)*$_2$ *(qu'il faisait beau)*$_1$ */ (être malade)*$_1$ */ (une bêtise)*$_1$[1].

[1]Dans ces phrases, $à\ N_2$ apparaît avant N_1, en accord avec le fait que les propriétés définitoires n'imposent pas d'ordre sur les compléments.

– Les propriétés définitoires occasionnelles servent à affiner la structuration en tables obtenue par les propriétés de base. Ainsi les verbes entrant dans le schéma $N_0\ V$ avec $N_0 = GN$ sont répartis en deux tables selon le trait plus ou moins humain du sujet. On peut donc inclure des traits sémantiques sur les $N_i = GN$ dans les propriétés définitoires. On peut aussi inclure des traits morphologiques soit sur le verbe (par exemple, la table 32RA a pour propriété définitoire $N_0\ V\ N_1$ avec $N_0 = GN$, $N_1 = GN$ et V est un verbe dérivé d'un adjectif, voir *assombrir* > *sombre*), soit sur un $N_i = GN$ (par exemple, la Table 32CV a pour propriété définitoire $N_0\ V\ N_1\ en\ N_2$ avec $N_0 = GN$, $N_1 = GN$, $N_2 = GN$ avec N_2 qui est dérivé d'un verbe, voir *caraméliser* > *caramel*). Enfin, les propriétés définitoires occasionnelles peuvent inclure une relation interdite ou autorisée entre deux phrases. Par exemple, la Table 32NM a pour propriété définitoire $N_0\ V\ N_1$ avec $N_0 = GN$, $N_1 = GN$ et où la forme passive est interdite (*Cette valise pèse 10 kilos, *10 kilos sont pesés par cette valise*). La Table 34L0 a pour propriété définitoire la relation de paraphrase entre $N_0\ V\ Loc\ N_1$ et, par abus de notation, $N_1\ V\ de\ N_0$ (*Des abeilles grouillent dans le jardin, Le jardin grouille d'abeilles*).

Les propriétés définitoires basiques et occasionnelles débouchent sur 61 tables. Les critères nombre et nature des compléments et réalisation des positions n'étant pas hiérarchisés, la structuration en tables peut être présentée de deux manières : (i) la présentation classique, où la réalisation des positions est le premier critère mis en avant, qui distingue les tables à complétive ou infinitive ($\exists\ N_i$ avec $N_i \neq GN$) regroupant les Tables 1 à 18 de (Gross, 1975), des tables sans complétive ni infinitive ($\forall\ N_i,\ N_i = GN$) regroupant les Tables 30 à 40 de (Boons *et al.*, 1976a; Boons *et al.*, 1976b; Guillet & Leclère, 1992) ; (ii) la présentation de (Leclère, 1990) où le nombre et la nature des compléments est le premier critère mis en avant.

Il nous reste à présenter l'aspect matriciel des tables. Un verbe (emploi de verbe) satisfaisant la propriété définitoire d'une table est l'en-tête lexicale d'une ligne de la table. Une table comporte plusieurs colonnes qui indiquent les propriétés respectées ou non par les en-têtes lexicales de la table. Les cases de la matrice relient les propriétés aux en-têtes lexicales avec des + et des -.

3 Introduction à DICOVALENCE

DICOVALENCE est un dictionnaire de valence verbale pour le français, héritier du lexique PROTON (Van den Eynde & Mertens, 2003). Il a été développé dans le cadre méthodologique de l'Approche Pronominale — cf. par exemple (Blanche-Benveniste *et al.*, 1984). Pour identifier la valence d'un prédicat (i.e. ses dépendants et leurs caractéristiques), l'Approche Pronominale exploite la relation qui existe entre les dépendants dits *lexicalisés* (réalisés sous forme de syntagmes) et les pronoms qui couvre « en intention » ces lexicalisations possibles. Les pronoms (et les paranoms, cf. ci-dessous), contrairement aux syntagmes, aux fonctions syntaxiques ou aux rôles thématiques, ont deux avantages majeurs :

– tout en étant des éléments de référence minimale, ils sont des éléments purement linguistiques, dénués des propriétés qui rendent difficile l'interprétation de la grammaticalité d'énoncés utilisant des dépendants syntagmatiques,
– ils sont en nombre restreint : leur inventaire est fini.

La valence peut donc être obtenue sans qu'il y ait besoin d'un travail d'interprétation, à l'aide d'une vérification systématique et exhaustive des combinaisons entre les différents pronoms et le prédicat verbal. Les pronoms retenus forment un ensemble plus large que ce que l'on désigne usuellement par le terme de « pronom » : il s'agit des pronoms clitiques, des pronoms personnels pleins et des pronoms dits *suspensifs* (qui regroupent ce que l'on appelle habituellement

pronoms interrogatifs et adverbes interrogatifs ou indéfinis, comme *à qui*, *quand*,. . .). Sont également pris en compte les *paranoms*, qui se distinguent des pronoms par leur modifiabilité (*rien* modifié dans *rien d'intéressant*) et l'impossibilité de reprise par un syntagme (**il ne trouve rien, les indices* vs. *il les trouve, les indices*).

Les combinaisons entre prédicats et pronoms induisent des paradigmes de portée globale. Certains correspondent à peu près aux traditionnelles fonctions syntaxiques (P0 = {*je*, *tu*, *il*, *elle*, *. . .*, *qui*, *. . .*} correspond à la fonction sujet (à l'exclusion du *il* impersonnel), P1 à la fonction objet direct, P2 à la fonction à-objet ou dative, etc.), d'autres permettent des distinctions plus fines que dans d'autres approches (PQ paradigme de quantité, PM paradigme de manière, etc.)[2].

DICOVALENCE proprement dit se présente comme une liste d'entrées correspondant chacune à un emploi d'un lemme verbal[3]. Sont tout d'abord donnés l'entrée et son type (prédicateur simple, verbe adjoint, verbe auxiliaire, verbe copule, verbe de dispositif, construction résultative[4]). Suivent alors les différents paradigmes qui dépendent du prédicateur (les termes de valences), avec pour chacun d'eux la liste des pronoms et paranoms qui peuvent en être la réalisation. Sont enfin indiquées certaines propriétés complémentaires, dont les passivations possibles (*passif être*, *se passif* et/ou *se faire passif*). Le tableau 1 présente l'entrée (unique) du verbe *supprimer* extraite de DICOVALENCE.

VAL$	supprimer: P0 P1
VERB$	SUPPRIMER/supprimer
VTYPE$	predicator simple
NUM$	80500
EX$	r : supprimer une loi / r : supprimer les obstacles
TR$	afschaffen, opheffen, intrekken, weghalen, weglaten, schrappen, doen verdwijnen
P0$	je, nous, on, qui, que, elle, il, ils, celui-ci, ceux-ci, ça
P1$	te, vous, qui, ceci, la, le, les, en Q, en, que, celui-ci, ceux-ci, ça, se$_{\text{réfl.}}$, l'un l'autre, se$_{\text{réc.}}$
RP$	passif être, se passif, se faire passif

TAB. 1 – Exemple d'entrée de DICOVALENCE (repris de (Van den Eynde & Mertens, 2006)).

4 Comparaison entre le *Lexique-Grammaire* et DICOVALENCE

4.1 Divergences fondamentales

Avant d'entrer dans des considérations scientifiques, signalons qu'une différence majeure entre le *Lexique-Grammaire* et DICOVALENCE réside dans le fait que DICOVALENCE est disponible librement dans son intégralité, alors que le *Lexique-Grammaire* n'est distribué que partiellement par ses dépositaires actuels[5].

Deux stratégies différentes ont été mises en œuvre dans le développement de ces ressources.

[2]Un même pronom peut appartenir à plusieurs paradigmes, p.ex. *nous* appartient à P0, P1 et P2.

[3]Il y a en moyenne 2,4 entrées par lemme.

[4]On se reportera à (Van den Eynde & Mertens, 2006) pour une description précise de ces termes.

[5]Environ 60% des données sont accessibles sur le site `http://infolingu.univ-mlv.fr/`. Pour les verbes pleins, seules 41 tables (sur 61) sont distribuées. Ceci empêche d'avoir une vision globale sur les entrées d'un verbe donné, certaines entrées pouvant faire partie de tables non distribuées.

DICOVALENCE se concentre volontairement sur les verbes les plus fréquents (3738 lemmes), et, pour ces verbes, sur leurs emplois les plus fréquents (8313 entrées). *A contrario*, le *Lexique-Grammaire* s'est lancé dans une quête sans limite d'exhaustivité, à l'intérêt discutable, aboutissant à 6500 lemmes décrits par 13 375 entrées. À titre d'illustration, la Table 31H des constructions de type $N_0^{hum}\ V$ comportait 129 entrées dans l'annexe de (Boons *et al.*, 1976a) ; cette même table, qui fait partie de celles qui sont distribuées librement, comporte aujourd'hui 626 verbes, dont certains lemmes peu usités tels que *bovaryser*, *calancher*, *se curedenter*, ou *faonner*. Au niveau des entrées, la même exhaustivité est recherchée : dans la table 9 (non disponible sauf dans (Gross, 1975)), qui, rappelons-le, a pour propriété définitoire $N_0\ V\ (Que\ P)_1\ à\ N_2$, inclut des verbes tels que *bourdonner*, *bramer*, *chuinter*, *coasser*, *couiner*, *croasser*, *crépiter*, ou *gargouiller* (ces entrées sont analysées par « fusion » : *bramer* $=$ *dire en bramant*). D'une manière plus générale, les verbes peu usités et les entrées douteuses (correspondant à des phrases traditionnellement préfixées par « ? ») sont considérés comme acceptables par le *Lexique-Grammaire*, alors que DICOVALENCE aura tendance à ne pas les conserver.

D'un point de vue méthodologique, le *Lexique-Grammaire* repose sur une structuration hiérarchique reflétée par une organisation en tables. À l'inverse, DICOVALENCE est un ensemble non structuré d'entrées. La structuration des entrées, que l'on peut formaliser par un graphe d'héritage, n'a aucune conséquence pour la réalisation d'analyseurs syntaxiques. Néanmoins, une telle structuration reflète des généralisations linguistiques pertinentes et facilite le développement et la maintenance de ressources lexicales. C'est ce qui a amené à la structuration en deux niveaux du Le*fff*, un niveau *intensionnel* structuré et un niveau *extensionnel* plat (Section 5.2)[6].

4.2 Où l'un empiète sur l'autre

Nous avons observé deux types d'empiètements entre les approches utilisées par le *Lexique-Grammaire* et par DICOVALENCE. D'une part, le *Lexique-Grammaire*, fait abondamment usage de l'approche pronominale, quoiqu'implicitement. Par exemple, le système Prép = Loc, utilisé pour les compléments prépositionnels introduits par les prépositions *à*, *de*, *dans*, *sur*, *sous*, etc., est entièrement fondé sur la pronominalisation de ces compléments par les pronoms *y*, *en*, *où*, *là*, *ici*, etc. De même, la distinction entre prépositions et complémenteurs pouvant apparaître devant les infinitives repose sur les propriétés de pronominalisation, tout comme le phénomène de chute de la préposition (et de « ce ») devant $Que\ P$. Ainsi, la table 8 regroupe des constructions du type $N_0\ V\ de\ (ce\ Que\ P)_1$ (*Luc doute de ce que Marie parte*) et du type $N_0\ V\ (Que\ P)_1$ (*Luc doute que Marie parte*), grâce à la pronominalisation en *en* et en *de cela* de la complétive, qu'elle soit ou non introduite par *de (ce)*.

D'autre part, DICOVALENCE ne se limite pas strictement à l'étude des réalisations pronominales des actants verbaux. Par exemple, le pronom *ça* peut pronominaliser une complétive à l'indicatif ou au subjonctif, mais aussi une infinitive, une concessive ou un interrogative indirecte. DICOVALENCE est donc obligé de spécifier les syntagmes dont *ça* peut être la pronominalisation. Ceci est fait par une mise entre parenthèses du type de syntagme correspondant : *ça(qpind)*, *ça(qpsubj)*, *ça(inf)*, etc. Le cas échéant, l'identifiant de réalisation syntagmatique est assorti d'un complémenteur : *ça(de_inf)*, *ça(à_inf)*, *ça(de ce qps)*, etc.

[6] Des considérations parallèles au niveau des grammaires ont induit le développement de la notion de *méta-grammaire*, description grammaticale structurée et factorisée, permettant la génération de grammaires classiques.

4.3 Où l'un apporte à l'autre

Le *Lexique-Grammaire* et DICOVALENCE sont deux ressources très riches mais incomplètes. Toutefois, elles peuvent mutuellement s'enrichir. Ainsi, DICOVALENCE comporte des informations précises sur les pronoms suspensifs (cf. plus haut), tandis que le *Lexique-Grammaire* ne prend pas en compte les interrogatives indirectes[7] et regroupent sous l'identifiant ADV les compléments pronominalisables en *ainsi, autant, (Prép) quand*, etc. A l'inverse, le *Lexique-Grammaire* comporte des informations plus précises que DICOVALENCE, en particulier :

– le système des noms parties du corps (N^{pc}), qui permet de rendre compte d'alternances telles que *Luc caresse les cheveux de Marie / Luc lui caresse les cheveux*,
– les compléments sous-catégorisés mais non pronominalisables (cf. Table 38PL : *couper le gâteau en quatre parts égales*, où le complément introduit par *en* n'est pas pronominalisable),
– certaines « restructurations », telles que *Luc copie les habitudes de Léa / Luc copie Léa dans ses habitudes*,
– certaines relations de dérivation morphologique : rappelons, par exemple, que la table 32RA regroupe les verbes morphologiquement dérivés d'adjectifs (*assombrir* > *sombre*).

Certaines propriétés ou constructions complexes, comme les différentes constructions pronominales, font l'objet de codages différents dans chacune des ressources. Dans un futur proche, nous chercherons à comprendre dans quelle mesure elles s'harmonisent et/ou se complètent.

En conclusion, ces deux ressources doivent impérativement être harmonisées et mutuellement enrichies pour obtenir une ressource lexicale complète du système verbal français. Nous nous sommes attaqués à cette tâche, avec l'objectif d'enrichir une ressource destinée au TAL. C'est ce que nous allons décrire dans la section suivante, consacrée au lexique Le*fff*.

5 Le Le*fff* : présentation et enrichissement

5.1 Historique

Le développement du Le*fff* a commencé en 2003, à partir du constat suivant : à cette époque, il n'existait pas de lexique syntaxique pour le français qui soit librement utilisable et dont la couverture soit importante. La construction d'un tel lexique a donc été initiée au sein du projet Atoll de l'INRIA par Lionel Clément, avec le double objectif qu'il soit adapté au TAL tout en restant linguistiquement pertinent.

Dans un premier temps, le Le*fff* s'est limité à un lexique morphologique des verbes du français, acquis automatiquement et validé manuellement selon une technique originale (Clément *et al.*, 2004; Sagot, 2005). C'est le Le*fff* 1, distribué depuis 2004. Dans un second temps, le Le*fff* a été étendu à l'ensemble des catégories, tout en devenant un lexique morphologique *et* syntaxique. L'extension à toutes les catégories a été faite manuellement pour les catégories fermées, et à l'aide du lexique morphologique français de Multext (Veronis, 1998) pour les noms, adjectifs et adverbes — lexique dont la libre exploitation nous a été autorisée explicitement par son principal auteur. Les informations syntaxiques ont été tout d'abord renseignées intégralement à la main, en profitant au mieux de l'architecture à deux niveaux du Le*fff* (cf. ci-dessous). Depuis, diverses techniques ont été utilisées pour étendre et corriger le Le*fff* : acquisition automatique (avec validation manuelle) d'entrées morphologiques et d'informations syntaxiques atomiques

[7] Voir cependant (Nakamura, 2006), qui a codé les interrogatives indirectes pour la Table 6 ($N_0\ V\ (Que\ P)_1$).

(Sagot, 2006; Sagot *et al.*, 2006), corrections et ajouts manuels ou guidés par des techniques automatiques telles la fouille d'erreurs dans les sorties d'analyseurs syntaxiques (Sagot & Villemonte de La Clergerie, 2006), et recherche de mots inconnus dans de grands corpus.

Le Le*fff* est donc aujourd'hui un lexique syntaxique à large couverture pour le français. Actuellement en version 2.5, il est entièrement téléchargeable sous sa forme extensionnelle, et sera prochainement téléchargeable également sous sa forme intensionnelle, sous une licence libre (LGPL-LR), sur le site internet `www.lefff.net`.

5.2 Modélisation des informations syntaxiques

Le Le*fff* repose sur une architecture à deux niveaux : (i) un *lexique intensionnel*, où l'information est factorisée au maximum, qui associe à chaque lemme une classe morphologique et une classe syntaxique ; c'est à ce niveau qu'est fait le travail de développement — (ii) un *lexique extensionnel*, obtenu à partir du lexique intensionnel par compilation, qui associe à chaque forme une structure représentant explicitement toutes les informations linguistiques associées ; c'est ce lexique qui est utilisé par les analyseurs. Ci-dessous une entrée intensionnelle pour le lemme *manger* et une entrée extensionnelle pour la forme fléchie *mange* :
manger v-er:std @verbe_transitif_direct,
mange v [pred='manger$_1$<Suj:sn|cln,Obj:(sn|cla)>', cat=v, @pers, @PS13s].

Au niveau intensionnel, les informations syntaxiques sont donc décrites à l'aide de classes syntaxiques, définies par héritage de propriétés syntaxiques atomiques, propriétés elles-mêmes définies de façon indépendante de la définition des classes. Au niveau extensionnel, le cadre de sous-catégorisation d'une forme donnée[8] est constitué d'une liste de *fonctions syntaxiques*, chacune indiquant les *réalisations* possibles de cette fonction ainsi que le caractère obligatoire ou non de sa réalisation (indiqué par des parenthèses). La structure syntaxique complète, outre le cadre, comporte le cas échéant des *macros* (introduites par « @ ») qui représentent de façon implicite des informations syntaxiques complémentaires (contrôle, attribution, (im)personnel,...).

Les fonctions syntaxiques ne sont utilisées ni dans le *Lexique-Grammaire* ni dans DICOVALENCE, mais les notions respectives de position et de paradigme s'en rapprochent. Elles sont définies dans le Le*fff* par des critères proches de ceux de DICOVALENCE, développés au cours de travaux de comparaison et fusion avec le *Lexique-Grammaire* et DICOVALENCE (en particulier sur les constructions impersonnelles, cf. (Sagot & Danlos, 2007)), mais également au cours de travaux non publiés en collaboration avec Claire Gardent. Ces critères reposent sur la substituabilité (en prenant en compte pronoms *et* syntagmes), sur le principe de réalisation unique d'une fonction syntaxique pour un prédicat donné, et sur l'identification de la fonction par un paradigme de pronoms (à l'exception des cas à partage d'arguments, c'est-à-dire les attributs)[9].

[8]Le passage de la forme intensionnelle à la forme extensionnelle est également un passage d'un lexique de lemmes à un lexique de formes. Ceci permet à certaines formes d'indiquer des modifications dans leur structure syntaxique par rapport à la structure par défaut correspondant à la classe syntaxique. Ainsi, bien que le conjugueur ne construise qu'une forme pour le participe passé d'un verbe passivable, le lexique extensionnel comportera deux entrées pour cette forme, l'une, active, avec le cadre par défaut, et l'autre, passive, dont les arguments syntaxiques seront caractéristiques de la construction passive.

[9]Actuellement, la liste de fonctions utilisées est la suivante : Suj (fonction sujet), Obj (fonction objet), Objà (fonction à-objet), Objde (fonction de-objet), Loc (fonction locative), Dloc (fonction délocative), Att (fonction attributive), Obl et Obl2 (fonctions obliques). Une terminologie assez traditionnelle a été préférée, pour des questions de lisibilité, à une terminologie plus algébrique comme utilisée dans DICOVALENCE. Ce qui ne signifie évidemment pas, par exemple, que toute réalisation d'un Objde comporte la préposition *de*, ni, à l'inverse, que tout complément

Les réalisations possibles, quant à elles, sont de trois types : *pronoms clitiques* (clitique nominatif (cln), accusatif (cla), datif (cld), génitif (en), locatif (y)[10]), *syntagme direct* (syntagme nominal (sn), adjectival (sa), infinitif (sinf), phrastique fini (scompl), interrogative indirecte (qcompl)) et *syntagme prépositionnel* (syntagme direct précédé d'une préposition, comme de-sn, à-sinf ou pour-sa ; à-scompl et de-scompl représentent les réalisations en *à/de ce que P*).

Le Le*fff* extensionnel est illustré dans le Tableau 2. On notera que les listes de fonctions syntaxiques dans les entrées active et passive de *mangé* sont présentées en ordre inverse. Ceci pour simuler d'une façon (trop ?) économique les rôles thématiques. Nous envisageons d'indiquer plus explicitement les rôles thématiques.

<table>
<tr><td>apprend</td><td>v</td><td>[pred='apprendre$_2$<Suj:sn|cln,Obj:(sn|cla|à-sinf|scompl|qcompl)>',</td></tr>
<tr><td></td><td></td><td>cat=v, @pers, @P13s] # Pierre apprend à conduire / la conduite</td></tr>
<tr><td>imagine</td><td>v</td><td>[pred='imaginer$_1$<Suj:sn|cln,Obj:(sn|cla), Att:(sn|sa|sinf|comme-sn|comme-sa)>',</td></tr>
<tr><td></td><td></td><td>cat=v, @pers, @PS13s] # Pierre imagine Marie nue / se dévêtir</td></tr>
<tr><td>mangé</td><td>v</td><td>[pred='manger$_1$<Suj:sn|cln,Obj:(sn|cla)>', cat=v, @active, @avoir, @Kms]</td></tr>
<tr><td>mangé</td><td>v</td><td>[pred='manger$_1$<Obl:(par-sn),Suj:sn|cln>', cat=v, @passive, @Kms]</td></tr>
</table>

TAB. 2 – Quelques entrées du Le*fff* extensionnel

5.3 Enrichir du Le*fff* à partir du *Lexique-Grammaire* et de DICOVALENCE

Le Le*fff* repose ainsi sur une architecture efficace et un format directement utilisable en TAL. De plus, ce format est en partie le résultat d'un consensus issu de travaux réalisés dans le cadre du projet ILF LexSynt. Nous avons donc commencé à enrichir le Le*fff* à partir du *Lexique-Grammaire* et de DICOVALENCE, ce qui demande une connaissance approfondie de ces deux ressources, étant donné les divergences entre les modèles lexicaux sous-jacents.

Il n'est pas possible de convertir directement au format Le*fff* les informations lexicales présentes dans le *Lexique-Grammaire*. Les travaux de (Gardent *et al.*, 2005) effectuent une telle conversion pour les tables distribuées (débouchant sur le lexique Synlex-Le*fff*), mais elle est indirecte, imparfaite, et nécessite une explicitation formalisée de données linguistiques sous-entendues dans le *Lexique-Grammaire*, apport qui n'est ni simple ni aisé. Nous avons donc préféré pour le moment nous focaliser sur certaines constructions, mal traitées dans le Le*fff*, et extraire du *Lexique-Grammaire* les informations lexicales pertinentes pour les y intégrer. Dans sa version actuelle (2.5), le Le*fff* a été amélioré de cette façon :

- pour les constructions impersonnelles verbales et adjectivales (Sagot & Danlos, 2007), extraites de l'outil ILIMP (Danlos, 2005) développé en partie à partir du *Lexique-Grammaire*,
- pour un certain nombre d'expressions verbales figées (Danlos *et al.*, 2006).

En revanche, la conversion de DICOVALENCE au format Le*fff* est plus directe. La convergence sur un nombre important de points entre les paradigmes de DICOVALENCE et les fonctions syntaxiques du Le*fff* rendent la correspondance relativement simple à implémenter[11]. Les pa-

introduit par *de* est la réalisation d'une fonction Objde. Pour une description plus précise, voir (Sagot & Danlos, 2007).

[10]On notera que le *se* réfléchi ou réciproque est considéré comme une réalisation de type cla ou cld selon les cas (*Les époux se disputent / Pierre se laisse cette possibilité*).

[11]Les correspondances, à quelques exceptions près (sujet des impersonnelles, par exemple), sont les suivantes : P0 → Suj, P1 → Obj, P2 → Objà, P3 → Objde, PL → Loc, PDL → Dloc, PMi → Att, PQ → Att (discutable), PP → Obl ou Obl2, PM ignoré (pour le moment).

radigmes de pronoms de DICOVALENCE peuvent également se convertir directement en listes de réalisations au sens du Le*fff*. Toutefois, le Le*fff* ne retranscrit pas ces paradigmes dans toute leur richesse, et de l'information est donc perdue. Elle pourrait ne pas l'être si l'on prenait également en compte les traits sémantiques que l'on peut extraire des paradigmes de pronoms (Mertens, comm. pers.). Mais pour l'instant, DICOVALENCE n'a été utilisé que pour procéder à une évaluation du Le*fff*, pas pour compléter et corriger ce dernier.

5.4 Exemple d'évaluation : les constructions verbales impersonnelles

L'évaluation d'un lexique comme le Le*fff* peut se faire naturellement via l'évaluation d'analyseurs qui reposent sur lui, travail que nous effectuerons dans le futur. Mais une évaluation directe par comparaison avec d'autres ressources est également riche d'enseignements. Pour illustrer le dialogue que nous avons instauré entre le *Lexique-Grammaire*, DICOVALENCE et le Le*fff*, nous avons procédé à une comparaison entre le Le*fff* et DICOVALENCE pour les entrées verbales impersonnelles, renseignées dans le Le*fff* à partir d'ILIMP, et donc indirectement à partir du *Lexique-Grammaire*. L'évaluation s'est faite sur les entrées défactorisées (une entrée comportant des disjonctions - sur les réalisations ou à cause de fonctions syntaxiques facultatives - est remplacée par un ensemble d'entrées). On peut alors comparer les cadres complets, ou se restreindre aux cadres fonctionnels (liste des fonctions syntaxiques réalisées).

Au niveau des cadres fonctionnels, les résultats sur les constructions verbales impersonnelles sont les suivants : 60 cadres présents à la fois dans DICOVALENCE et dans le Le*fff*, 160 cadres (tous corrects) présents seulement dans le Le*fff*, et 19 cadres présents uniquement dans DICOVALENCE (certains d'entre eux nous ont semblé inacceptables, certains autres sont la conséquence de difficultés de conversion entre DICOVALENCE et le modèle lexical du Le*fff*). Le Le*fff* est donc désormais couvrant et précis sur les impersonnelles.

6 Conclusion

L'amélioration du Le*fff* à partir du *Lexique-Grammaire* et de DICOVALENCE est donc en cours dans le but d'obtenir une ressource de référence pour le TAL[12]. En ce qui concerne le *Lexique-Grammaire*, la pertinence et l'utilité de ces travaux reste limitée, compte tenu de la disponibilité restreinte des tables. Ceci n'est pas le cas de DICOVALENCE et nous comptons effectuer à partir de cette ressource, en collaboration avec Piet Mertens, un travail sur les constructions verbales pronominales pour augmenter la qualité du Le*fff* sur ce point.

En amont des données lexicales proprement dites, le modèle lexical du Le*fff* est à améliorer : certains cas verbaux non triviaux sont à revoir (en particulier, les modaux et *verbes adjoints* (selon la terminologie de DICOVALENCE), qui correspondent respectivement aux verbes de la Table 1 et aux verbes de perception de la Table 6 du *Lexique-Grammaire*), le modèle utilisé pour les constructions figées est trop simpliste, certaines fonctions syntaxiques sont probablement à découpler (distinguer Att d'un pseudo-objet de type PQ) ou à ajouter (cf. PM), et certaines réalisations, comme les concessives, sont à mieux prendre en compte.

Enfin, nous allons mettre l'accent sur différents types d'interface de visualisation et d'édition du Le*fff*, voire de comparaison avec d'autres ressources converties au format Le*fff*.

[12]Des travaux de comparaison avec Synlex-Le*fff* sont aussi en cours, en collaboration avec Claire Gardent.

Références

BLANCHE-BENVENISTE C., DELOFEU J., STEFANINI J. & EYNDE K. V. D. (1984). *Pronom et syntaxe. L'approche pronominale et son application au français*. Paris : SELAF.

BOONS J.-P., GUILLET A. & LECLÈRE C. (1976a). *La structure des phrases simples en français, Classes de constructions transitives*. Rapport interne, LADL, CNRS, Paris 7.

BOONS J.-P., GUILLET A. & LECLÈRE C. (1976b). *La structure des phrases simples en français, Constructions intransitives*. Genève : Droz.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proc. of LREC'04*, p. 1841–1844, Lisbon, Portugal.

DANLOS L. (2005). ILIMP : Outil pour repérer les occurences du pronom impersonnel *il*. In *Actes de TALN 2005*, Dourdan, France.

DANLOS L., SAGOT B. & SALMON-ALT S. (2006). French frozen verbal expressions : from lexicon-grammar to NLP applications. In *Actes du colloque sur le lexique et la grammaire 2006*, Palerme, Italie.

GARDENT C., GUILLAUME B., PERRIER G. & FALK I. (2005). Maurice Gross' grammar lexicon and natural language processing. In *Proc. of the 2nd LTC*, Poznań, Poland.

GROSS M. (1975). *Méthodes en syntaxe*. Paris : Hermann.

GUILLET A. & LECLÈRE C. (1992). *La structure des phrases simples en français : Les constructions transitives locatives*. Genève : Droz.

LECLÈRE C. (1990). Organisation du lexique-grammaire des verbes français. *Langue Française*, **87**.

NAKAMURA T. (2006). *Lexique et grammaire des interrogatives partielles en français : étude des verbes à une complétive directe*. PhD thesis, Université de Marne-la-Vallée.

SAGOT B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Proc. of TSD 2005 (LNAI 3658, ©Springer-Verlag)*, Karlovy Vary, Czech Rep.

SAGOT B. (2006). *Analyse automatique du français : lexiques, formalismes, analyseurs*. PhD thesis, Université Paris 7.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE E. & BOULLIER P. (2006). The Le*fff* 2 syntactic lexicon for French : architecture, acquisition, use. In *Proc. of LREC'06*.

SAGOT B. & DANLOS L. (2007). Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles et expressions verbales figées. *Cahiers du Cental*. to appear.

SAGOT B. & VILLEMONTE DE LA CLERGERIE E. (2006). Error mining in parsing results. In *Proc. of ACL 2006*, p. 329–336, Sydney, Australia.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, **13**, 63–104.

VAN DEN EYNDE K. & MERTENS P. (2006). Le dictionnaire de valence DICOVALENCE : manuel d'utilisation. http ://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf.

VERONIS J. (1998). *Multext-Lexicons, A set of Electronic Lexicons for European Languages*. Rapport interne.